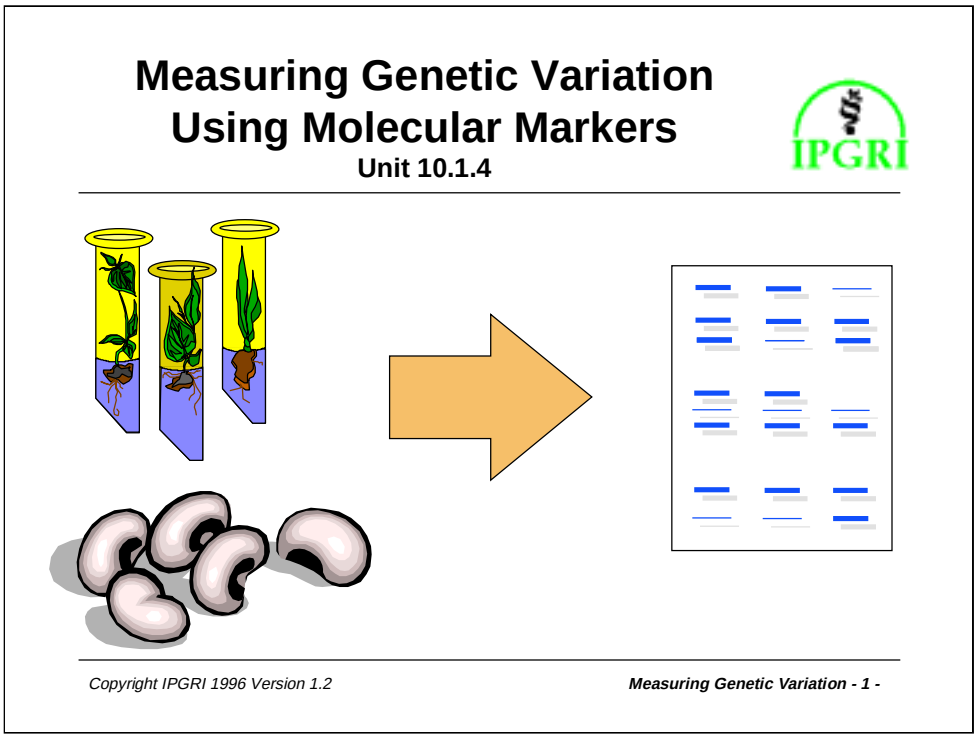


Unit 10.1.4

Measuring Genetic Variation Using Molecular Markers

Version 1.2

Brian Ford-Lloyd, University of Birmingham

Kevin Painting, IPGRI, Rome

Acknowlegements

- Authors
- Reviewers
- Figures
- Illustrations and graphics

Authors

Brian Ford-Lloyd, University of Birmingham
Kevin Painting, IPGRI, Rome

Reviewers

The authors would like to thank the following people for their critical review of the manuscript and their useful suggestions: Geo Coppens, Ian Dawson, Jan Engels, Hareya Fassil, Thomas Gass, Christopher Gooding, Luigi Guarino, Devra Jarvis, Elizabeth Jones, and Lyndsey Withers.

Figures

Model of insulin (page 8) appears courtesy of Swiss 3D Image, Geneva University Hospital and University of Geneva, Geneva, Switzerland.

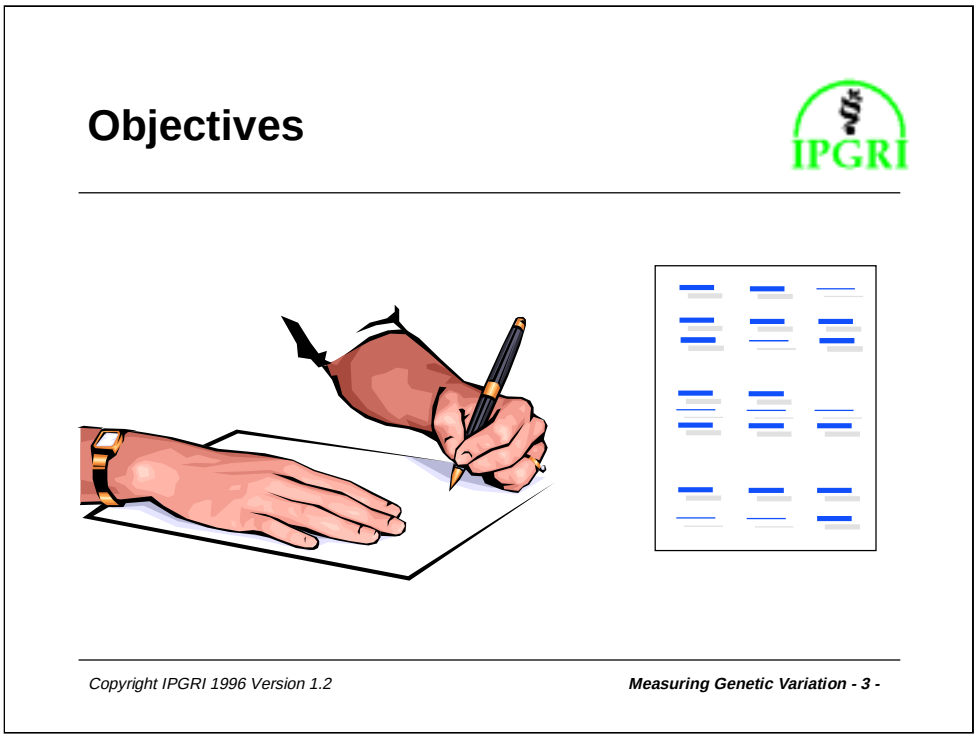
Model of DNA (page 18) appears courtesy of the Image Library of Biological Macromolecules, Jena, Germany.

D. Marshall and B. Parsons, University of Birmingham for RFLP (page 27) and RAPD (page 65) gels.

J. Xu, G. Bryan, John Innes International for AFLP gel (page 50).

Illustrations and Graphics

Illustrations by Kevin Painting and Lyndsey Withers, IPGRI, and New Vision Technologies Inc.



Objectives

Upon completion of the unit the trainee will be able to:

• define molecular markers and to give examples of situations where they can be useful in plant genetic resources work

• review protein-based molecular markers, summarising the protocols used, providing an appreciation of technical difficulty and highlighting the advantages and / or limitations of their use

• review DNA-based molecular markers with particular reference to RFLP and PCR-based methods, summarising the protocols used, providing an appreciation of technical difficulty and cost and highlighting the advantages and / or limitations of their use

• summarise the different electrophoretic techniques used to characterise molecular markers

• outline the theoretical and practical considerations to be made when choosing a molecular marker or technique

Topics

- Introduction
- Protein-based markers
- DNA basics
- RFLP-based markers
- PCR
- RAPD
- Sequence-tagged sites
- AFLP
- Electrophoretic techniques
- Considerations when choosing a technique
- Practical applications

Topic 1 - Introduction

- Environmental factors can influence morphological characters
- Molecular markers focus on the genetic material or the variation controlled by genes
- Molecular markers can be used for:
  - measuring several genetic diversity indicators
  - gene mapping
  - selection of germplasm in breeding programmes

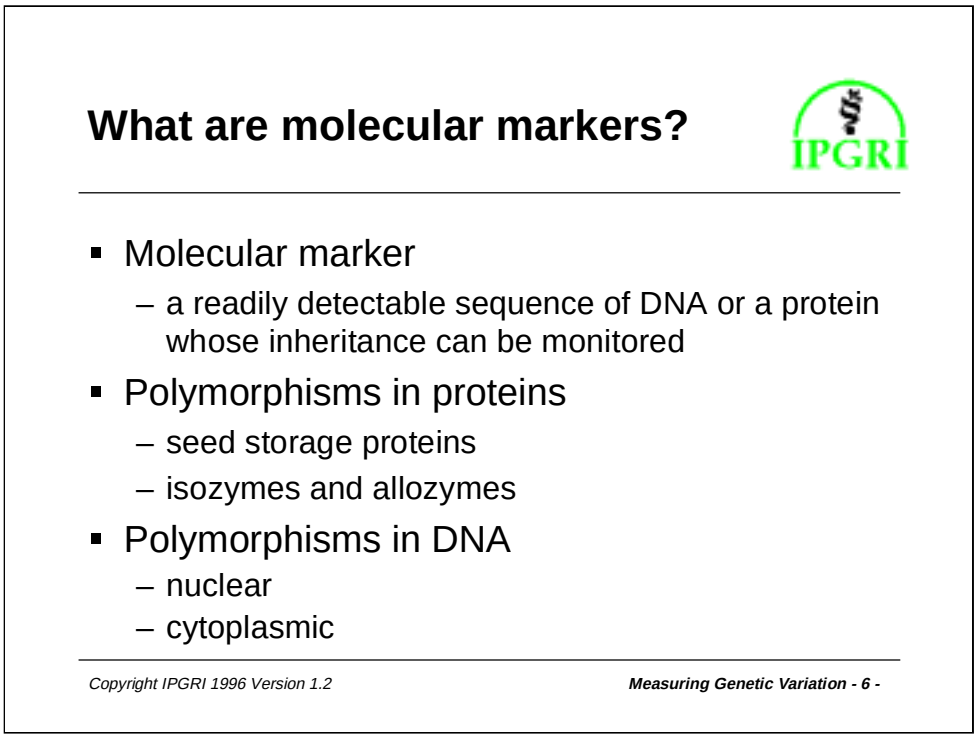
In plant genetic resources conservation, the ultimate aim is to conserve, as far as possible, the broad genetic diversity which is found in the target species. This aim recognises that genetic diversity is a critical component of biodiversity and environmental issues and that the genetic resources themselves are a rich potential source of useful genetic traits.

Classical methods of estimating the genetic diversity or relatedness among groups of plants have relied upon morphological characters. However, these characters can be influenced by environmental factors. Molecular and biochemical markers avoid many of the complications of environmental effects acting upon characters by looking directly at variation controlled by genes or by looking at the genetic material itself. Molecular markers represent a powerful and potentially rapid method for characterising diversity for *in situ* and *ex situ* conservation.

With molecular markers, direct and accurate measurements of the following genetic diversity indicators can be made:

• Gene diversity statistics of Nei (1973)
• Population subdivision of Wright (1965) and Lynch (1990)
• Heterozygosity, effective population size, allele frequency
• Estimating similarity and distance (genetic distance), using cluster analysis.

Molecular markers are very useful in the construction of genetic maps and for the screening and selection of germplasm within breeding programmes.



A molecular marker is a sequence of DNA or a protein which can be readily detected and whose inheritance can be monitored. It is the variation in, or *polymorphism* of, molecular markers which can be used in genetic diversity studies.

Polymorphisms in proteins

• **seed storage proteins**

• **isozymes and allozymes** are different molecular forms of an enzyme sharing a catalytic activity. **Allozymes** are different molecular forms of an enzyme coded by different alleles at one gene locus. **Isozymes** are different molecular forms of an enzyme coded by more than one gene locus. However, the term 'isozymes' is freely used for both situations.

Polymorphisms in DNA

DNA can be compared in many different ways in order to identify polymorphisms. These polymorphisms will be represented by differences in the DNA sequences.

Nuclear and cytoplasmic DNA (namely, chloroplast DNA [cpDNA] and mitochondrial DNA [mtDNA]) can be studied for polymorphisms. Both chloroplast and mitochondrial genomes are considerably smaller than the nuclear genome in terms of base composition, but both occur in multiple copies in each cell. They are generally transmitted maternally except in gymnosperms.

Desirable properties of molecular markers

- Polymorphic
- Codominant inheritance
- Occurs thoughout the genome
- Easy, fast and inexpensive to detect
- Reproducible/transferable

No single marker meets all needs

Desirable Properties

• **Polymorphic**. The marker *must* be polymorphic as it is the polymorphism itself which is measured for genetic diversity studies. However, the level of polymorphism detected can vary depending on the method used to measure it.

• **Codominant inheritance.** The different forms of the marker should be detectable in diploid organisms to allow the discrimination of homo- and heterozygotes.

• **Occurs throughout the genome.** Except where a marker is a specific locus, the marker should be evenly and frequently distributed throughout the genome.

• **Easy, fast** and **inexpensive** to detect.

• **Reproducible** within and between laboratories.

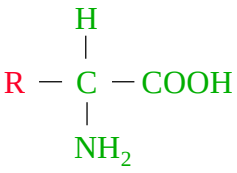
Unfortunately, there is no single molecular marker which meets all these needs. Some will be useful because they are co-dominant and can be used in breeding, others can detect variation in a wide range of different species and are therefore widely applicable, and others will be much more reproducible and transferable than others.

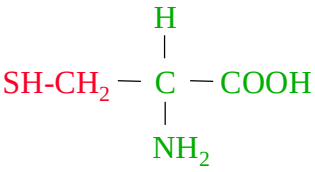
Topic 2 - Protein-based markers

- An introduction to proteins
- Seed storage proteins
- Isozymes and allozymes

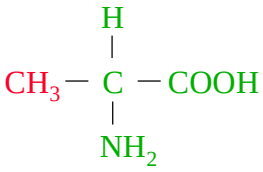
Protein structure - the basics

- Amino acids
  - the building blocks





cysteine



alanine

The building blocks of proteins are amino acids. These are bifunctional organic compounds which contain a basic amino group ($-NH_2$) and an acidic carboxyl group (-COOH). In proteins, twenty different amino acids are commonly found which vary in properties according to the nature of the R-group. For instance, in alanine the R-group is ($-CH_3$); in cysteine it is ($-CH_2-SH$).

In proteins, amino acids are joined together in chains by peptide (amide) bonds which form the backbone of the molecule. A peptide bond is formed between a basic amino group ($-NH_2$) on one amino acid and an acidic carboxyl group (-COOH) on another. The term 'polypeptide' simply refers to a long chain of amino acids. A 'protein' can be made up of one or more separate polypeptides. Clearly, the properties of polypeptides and proteins depend on their amino acid composition.

Protein structure - cont[d]

- Protein structure
  - Primary structure - polypeptide backbone
  - Secondary structure - local hydrogen bonds
  - Tertiary structure - bonding involving R groups
  - Quaternary structure - polypeptide:polypeptide interaction

Proteins have a primary, a secondary, a tertiary and a quaternary structure.

• The **primary structure** is the order of amino acids on the polypeptide chain - the polypeptide backbone.

• The **secondary structure** is the result of local hydrogen bonds along the polypeptide backbone. As a result, characteristic shapes are formed by folding of the polypeptide chains. This gives the protein strength and flexibility. Common structures found are:

- *alpha-helix* - caused by hydrogen-bonding within the polypeptide chain, e.g. muscle proteins.
- *beta-pleated sheet* - caused by hydrogen-bonding between adjacent polypeptide chains, e.g. silk fibroin.

• The **tertiary structure** results from interactions between the R-groups in a polypeptide, such as non-covalent bonds (hydrogen bonds, ionic bonds, hydrophobic interactions) and weak, covalent bonds (disulphide bonds between cysteine residues).

• The **quaternary structure** results from interactions between two or more polypeptide chains to form dimers, trimers, tetramers, etc. These are held together by hydrogen bonds, ionic bonds and, less commonly, hydrophobic interfaces and interchain disulphide bonds.

Protein structure - cont^{d}

- Effects of the environment
- Denaturation
- Diversity of protein function
  - facilitated by complexities of protein structure

The three-dimensional structure that proteins have is a direct result of interactions with their environment. For instance, in aqueous environments the hydrophobic R-groups are positioned toward the interior of the protein. Changes in temperature or pH can interfere with the non-covalent bondings causing disruption in the three-dimensional structure and loss of activity. This process is called *denaturation*. Denatured proteins can also clump together to become insoluble in a process called *coagulation*.

Diversity of protein function - *facilitated by complexities of protein structure*

- **Structural** - e.g. collagen, muscle fibres.
- **Storage** - e.g. wheat gliadins, barley hordeins.
- **Enzyme** - e.g. hydrolases, transferases, isomerases, etc.
- **Transport** - e.g. oxygen transfer with haemoglobin.
- **Messengers** - e.g. insulin and certain other hormones.
- **Antibodies** - proteins which bind to specific foreign particles.
- **Regulation** - e.g. proteins involved in regulating DNA synthesis.

Seed storage proteins

- Why use proteins from seeds?
  - seeds are a rich source of proteins
  - proteins are available in sufficient quantity
  - seeds are a well-defined stage in development
- Methodology
  - extract proteins
  - separate proteins by electrophoresis
  - visualise proteins on gel by staining
  - analyse banding patterns

Seed storage proteins of many species have been used in diversity studies since the 1960s. Seeds have been used in preference to other plant structures such as leaves because:

- seeds are a rich source of proteins which are generally stable
- the storage proteins are present in sufficient quantity for analysis
- seeds represent a well-defined stage in the plant's development.

Methodology

• Extract the desired seed storage proteins using an appropriate extracting medium and pH (alcohol / water based; dilute acid or alkali).

• Separate these proteins according to their size and net charge by gel electrophoresis, usually polyacrylamide gel electrophoresis (PAGE). For higher resolution, isoelectric focusing is used.

• Visualise the proteins on the gel using a general protein stain such as Coomassie blue or Imido Black.

• Analyse the banding patterns obtained.

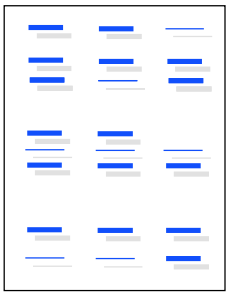
Rationale for analysis

The rationale behind the analysis is that:

- each band in the profile represents a direct gene product
- profiles are a measure of the genetic similarities / differences among the plants in the study.

Considerations and applications

- Considerations
  - co-migration
  - complexity of banding patterns
  - intraspecific variation
- Application of the technique
  - e.g. wheat, barley



Considerations

• **Co-migration: same mobility, same protein?** An assumption commonly made when comparing profiles is that proteins which share the same mobility and intensity in a gel are *homologous* proteins, that is, that they are products of the same gene(s). This is a questionable assumption, especially if only one set of electrophoresis conditions is used. Further genetic tests are the only way of confirming this assumption although repeating the electrophoresis under different conditions (e.g. gel concentration, pH) can give the assumption more credibility. Co-migration also occurs with certain DNA molecular markers.

• **Complexity of banding patterns.** Twenty or more individual bands can be resolved in a typical seed protein profile. However, the greater the number of bands, the greater the problems associated with co-migration.

• **Intraspecific variation** can be quite high even among closely related species. This variation must be established before making inter-specific comparisons. Similarly, intrapopulation variation should be measured even if studies to date have found this to be low or nonexistent.

Applications of the technique

Examples: gliadins from wheat, hordeins from barley.

Storage proteins can provide useful genetic profiles at low cost for the identification of genotypes or varieties. For example, in wheat they can provide useful markers for breadmaking qualities. However, owing to the advantages of enzyme electrophoresis, both the popularity and application of seed protein electrophoresis have declined.

Isozymes and allozymes

- Multiple forms of the same enzyme
  - allozyme: one enzyme, one gene locus
  - isozyme: one enzyme, more than one gene locus
- Methodology
  - macerate tissue
  - separate enzymes by electrophoresis
  - locate enzymes by histochemical staining
  - analyse banding patterns

Enzyme electrophoresis has been very useful in genetic diversity studies as it can directly reveal genetic polymorphism through demonstrating multiple forms of a specific enzyme. The multiple forms of an enzyme are of two classes:

• **Allozymes** - the enzyme is coded for by different alleles at *one* gene locus.
• **Isozymes** - the enzyme is coded for by alleles at *more than one* gene locus.

In common usage the term 'isozymes' refers to both classes.

Methodology

• Macerate appropriate plant tissue (e.g. young leaves) in cold buffer.

• Separate enzymes from the tissue extract using starch or polyacrylamide gel electrophoresis.

• Locate isozymes by histochemical staining - incubate gel with an enzyme-specific substrate which gives a coloured product when metabolised thereby revealing location of the isozyme on the gel.

• Analyse banding patterns.

Examples of enzymes studied: esterase, acid phosphatase, catalase, phosphoglucoisomerase.

Interpretation of banding patterns

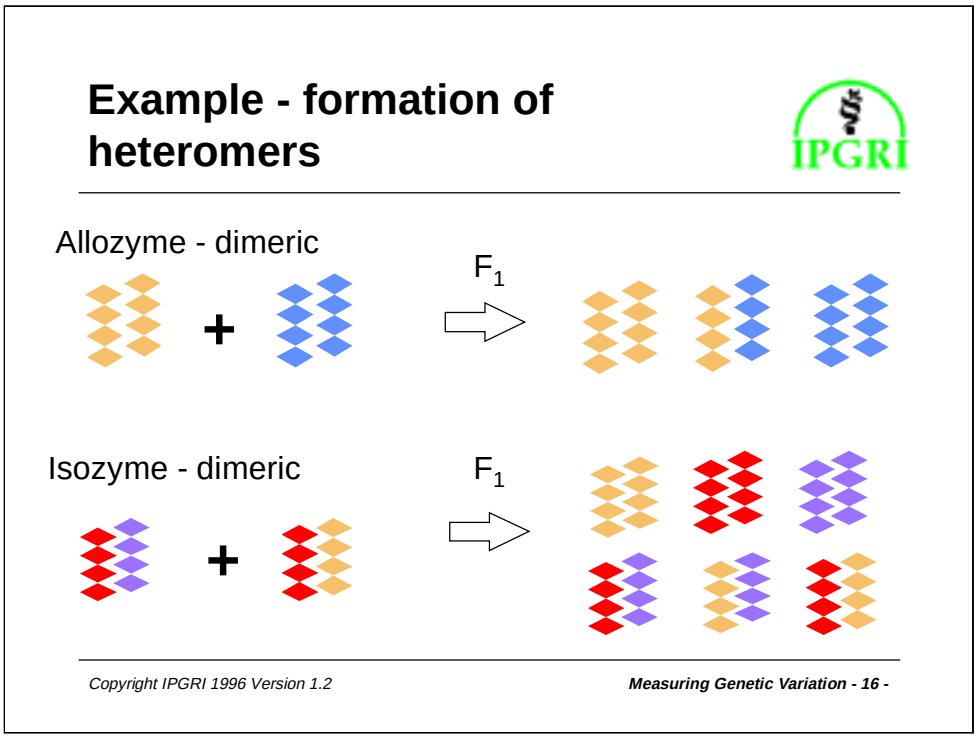
- Homozygous or heterozygous ?
- Quaternary structure of the enzyme
- Number of gene loci
- Number of alleles
- Genetic analysis sometimes necessary

The principal considerations here are:

- whether the organism is homozygous or heterozygous at the gene loci
- the quaternery structure of the enzymes (monomeric, dimeric, etc.)
- the number of gene loci
- the number of alleles per locus.

Allozymes are controlled by codominant alleles which means that it is possible to distinguish between homozygotes and heterozygotes. For monomeric enzymes (i.e. consisting of a single polypeptide), plants homozygous for that locus will produce one band whereas heterozygous individuals will produce two. For dimeric enzymes (i.e. consisting of two polypeptides), plants homozygous for that locus will produce one band whereas heterozygous individuals will produce three owing to random association of the polypetides. With tetrameric enzymes, heterozygous individuals will have five bands. For multimeric enzymes where the polypeptides are specified by different loci, the formation of isozymic heteromers can complicate the banding patterns considerably. This is shown in the next slide.

In view of these complexities and the importance of correctly interpreting banding patterns, genetic analysis is desirable and often necessary. For instance, in self-incompatible plants, this can be achieved by crosses with individuals with known banding patterns followed by analysis of the progeny (F_1, F_2 and backcross).



These examples show the potential behaviour of dimeric enzymes in crosses.

In the first example, both parents are homozygous for different alleles at the same locus. In such a cross, all F_1 progeny are heterozygous but because of random association of the polypetides, three combinations are possible and therefore three bands will be resolved upon gel electrophoresis and staining.

In the second example, the two polypeptides of the enzyme are coded by separate loci. However, both parents have the same allele at one of these loci (shown in red). In such a cross, random association of the polypeptides can lead to a total of six heteromers which can be resolved upon gel electrophoresis and staining.

Applications

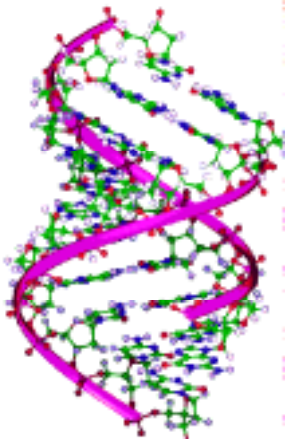
- Robust, reproducible method for:
  - characterising/identifying genotypes
  - studying population genetics
  - examining geographical patterns of variation

- Limited number of enzymes available

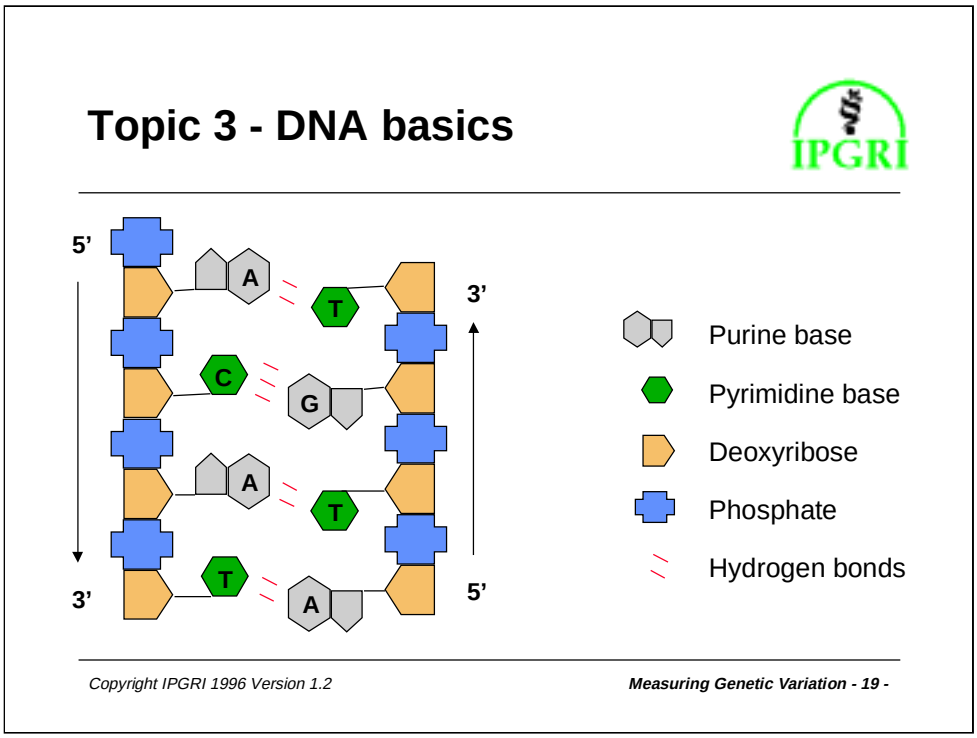
Isozyme polymorphism has been used for characterising and identifying genotypes and varieties of crop plants, for studying population genetics and for examining geographical patterns of variation.

Well over 30 enzyme systems have been used in plants, and for some crop plants the genes involved have been mapped. However, there is a definite limit to the number of enzymes available for study (because of availability of ways of detection), and hence the proportion of any plant genome which can be accessed.

DNA-based markers



- DNA basics
- RFLP methods
- PCR-based methods
- Sequence-tagged sites
- AFLP
- Electrophoretic techniques



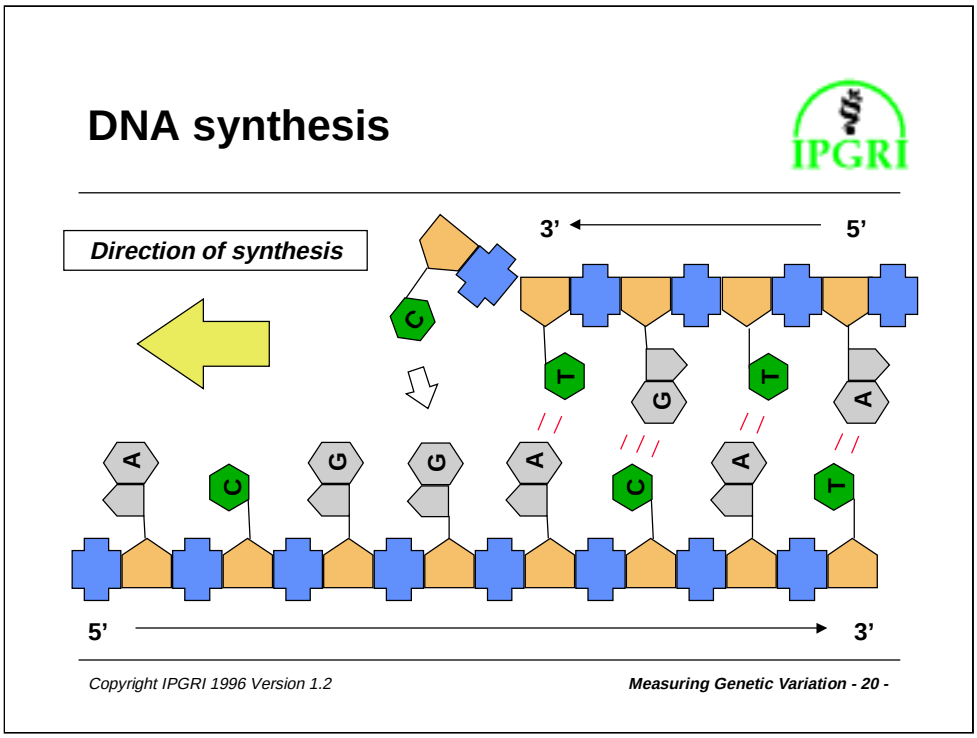
DNA structure

The building blocks of DNA (and RNA) are nucleotides. A nucleotide consists of:

- a **pentose sugar** - in DNA it is deoxyribose, in RNA it is ribose
- a **phosphate group**
- a **nitrogenous base**
    - purine bases: adenine (A), guanine (G).
    - pyrimidine bases: cytosine (C), thymine (T). In RNA, uracil (U) replaces thymine.

In DNA, the nucleotide building blocks are assembled to form a double helix. This consists of two strands, each with a sugar-phosphate backbone, held together by weak hydrogen bonding between the bases adenine-thymine (two hydrogen bonds) and cytosine-guanine (three hydrogen bonds). The strands run *anti-parallel*, that is they run in opposite directions with respect to the 5' and 3' ends of the molecules.

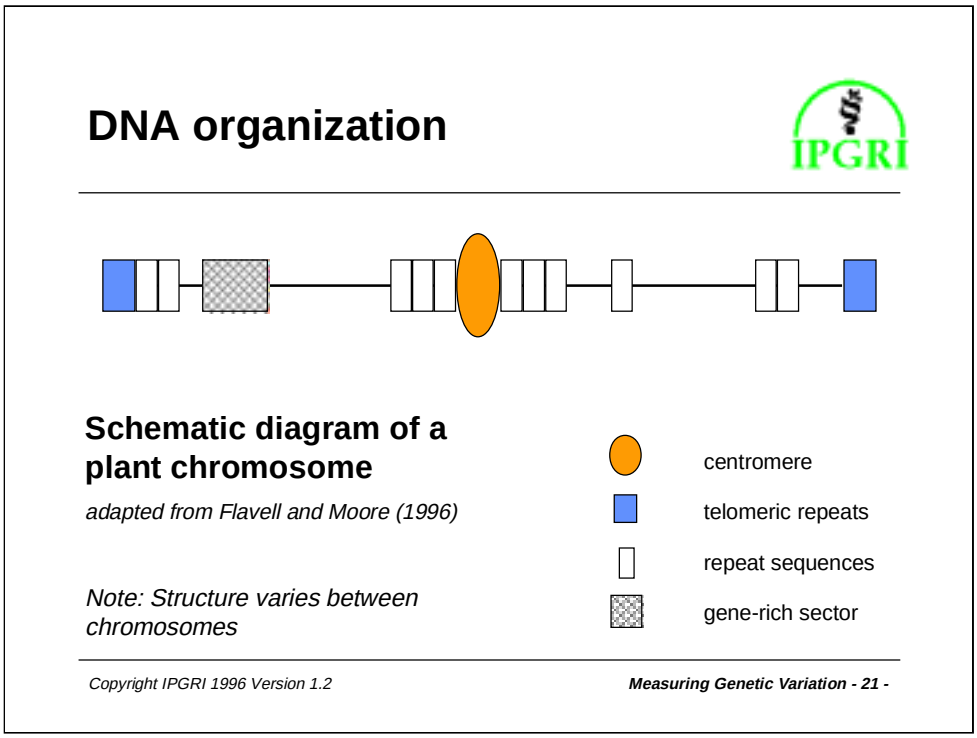
The DNA sequence is the order of the nucleotide bases along a DNA strand from the 5' to 3' end; this sequence specifies precise genetic instructions for the organism and is represented using single-letter codes, e.g. GATCGG and so on.



DNA synthesis

In cellular DNA replication, the DNA is first unzipped and denatured into single strands by a number of enzymes. The RNA polymerase next synthesises a short stretch of RNA complementary to one of the DNA strands at a particular site (the replication start site). This short section of RNA acts as a *primer* for the start of DNA replication. The DNA polymerase now starts to elongate the primer at the 3' end by adding nucleotides complementary to those on the template DNA strand. As nucleotides are always added at the 3' end, DNA synthesis occurs from a 5' to 3' direction.

Note: Replication occurs on *both* strands of the parent DNA.



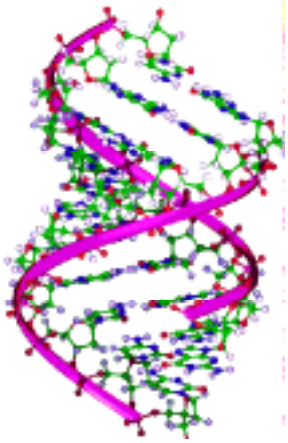
In order to protect the large amount of DNA and to regulate its expression, DNA is packaged into specific structures at the cellular level. In eukaryotes, DNA is condensed with histone and non-histone proteins (and some RNA) into thread-like structures called *chromosomes* which are found in the nucleus. Smaller amounts of DNA are found in the cytoplasm - in the chloroplasts [cpDNA] and in the mitochondria [mtDNA]. In prokaryotes, DNA is found supercoiled in a compact structure called the *nucleoid* which also contains proteins and some RNA.

In eukaryotes, the number of chromosomes varies between species and occasionally within species. At the subchromosomal level, several types of organization are observed. These can be summarised as follows:

- **Gene-rich sectors.** In large genomes, genes are found clustered in gene-rich sectors especially in regions close to the telomeres. In a number of cases, it is significant that the order of genes in a sector is conserved between species ('gene synteny'). Genes in a gene-rich sector are interspersed with short repeat sequences, often transposable elements.

- **Tandem repeats**. Multiple repeats of essentially the same sequence are found clustered at many locations, especially around the centromeres, telomeres and interstitial locations. These arrays can consist of up to millions of repeat units. Tandem repeats vary according to the size and sequence of the repeat unit, the number of repeats found and their distribution throughout the genome. They have therefore received considerable attention as molecular markers.

DNA-based markers



- DNA basics
- **RFLP methods**
- PCR based methods
- Sequence-tagged sites
- AFLP
- Electrophoretic techniques

Topic 4 - RFLP based markers

- RFLP examines differences in size of specific DNA restriction fragments
- Usually performed on total cellular DNA
- Requires pure, high molecular weight DNA

Restriction fragment length polymorphism (RFLP) analysis was one of the first techniques to be used widely to detect variation at the sequence level. It examines the variation in size of specific DNA fragments following digestion with restriction enzymes. RFLP analysis (and variations in the technique) have the same basic steps, namely:

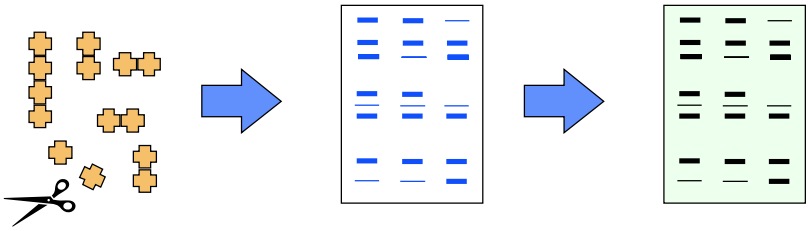
- Isolation of DNA
- Cutting DNA into smaller fragments using restriction enzymes
- Separation of DNA fragments by gel electrophoresis
- Transferring DNA fragments to a filter
- Visualising specific DNA fragments using a radioactively labelled probe
- Analysis of results.

Isolation of DNA

Usually *total* DNA is extracted from plant cells. However, chloroplast DNA and mitochondrial DNA can be isolated separately if required for individual analysis. The DNA must be very pure and of high molecular weight. This is sometimes difficult because of (a) degradation of DNA - by nucleases present in the plant cells; (b) isolation of polysaccharides which make handling DNA difficult; (c) isolation of secondary plant metabolites which damage the DNA or inhibit the restriction enzymes used. With PCR-based methods (see later), the requirement for clean, high molecular weight DNA is not so stringent.

RFLP methodology (cont[d])

- Cutting DNA into smaller fragments
- Separating fragments by gel electrophoresis
- Transferring DNA fragments to a filter



Cutting DNA into smaller fragments using restriction enzymes

Extracted DNA is digested with specific, carefully chosen restriction enzymes. Restriction enzymes recognise and cut specific palindromic DNA sequences, usually consisting of four to six bases. Each restriction enzyme, under the appropriate conditions, will recognise and cut DNA in a predictable way resulting in a reproducible set of DNA fragments ("restriction fragments") of different lengths. A large number of restriction enzymes are commercially available, each recognising and cutting specific DNA sequences.

Separation of DNA fragments by gel electrophoresis

The millions of restriction fragments produced are commonly separated by electrophoresis on agarose gels. Since the fragments would be seen as a continuous 'smear' if stained with ethidium bromide, staining alone cannot detect the polymorphisms. It is necessary therefore to detect specific fragments using hybridisation methods.

Transfering DNA fragments to a filter

The restriction fragments are transferred to a nylon or nitrocellulose membrane filter such that the fragments retain the same pattern on the membrane as on the gel. The process is called "Southern blotting", named after the person who invented the technique.

RFLP methodology (cont[d])

- Visualise DNA fragments
  - radioactive probes
  - non-radioactive probes
- Analysis of results
  - bands scored for presence/absence
  - differences in band patterns reflect genetic differences

The choice of probe/restriction enzyme is crucial

Visualising specific DNA fragments using a radioactively labelled probe

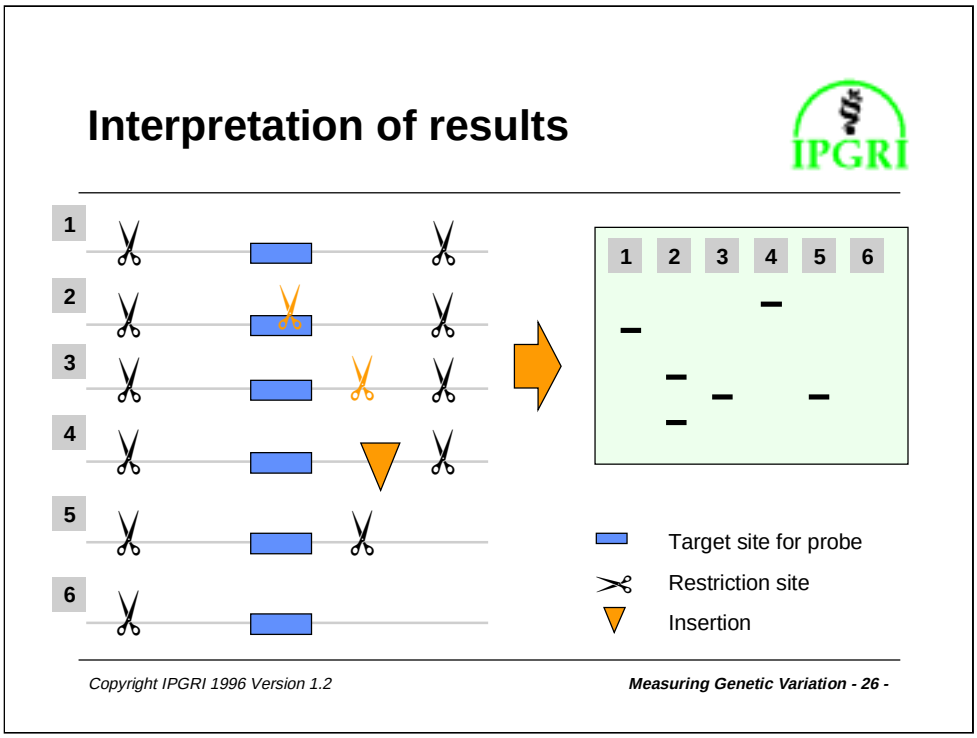
The relative positions of the desired RFLP marker can be visualised by firstly incubating the filter with a radioactively labelled DNA probe containing the RFLP marker sequence and subsequently detecting the hybridised fragments using autoradiography.

Alternatively, non-radioactive probes can be used. In this case, the hybridised fragments are then detected by non-radioactive assays involving chemiluminescence or colorigenic detection.

Analysis of results

The result is ideally a series of bands on a gel which can then be scored for presence or absence of particular bands. Differences between genotypes are usually visualised as an altered pattern of DNA restriction fragments. This may result from point mutations creating or destroying restriction sites, or because of reorganisations of blocks of DNA, such as deletions or insertions, between restriction sites (see next slide).

It is clear that choice of the DNA probe/restriction enzyme combination is crucial in the discriminating power of RFLP technology. The choice of DNA probe will be discussed later.



This slide (after Ferreira and Grattapaglia, 1996), shows the different mutational events responsible for the polymorphisms detected by RFLP analysis.

Track 1 - 'wild type'.

Track 2 - A mutation creating a new restriction site occurs within the target region. Two smaller bands are therefore detected on autoradiography.

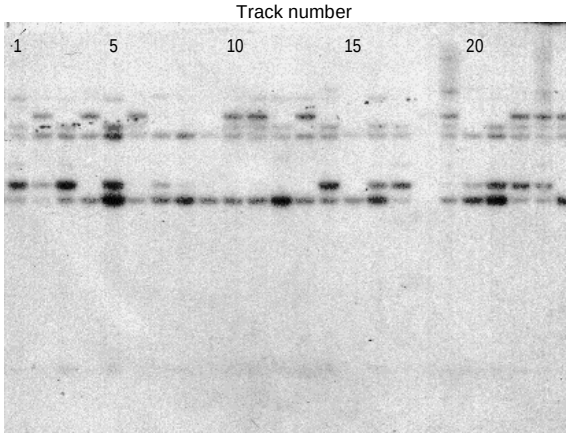
Track 3 - A mutation creating a new restriction site occurs between the flanking restriction sites, creating a smaller restriction fragment.

Track 4 - Insertion of a DNA sequence occurs between the flanking restriction sites, creating a larger restriction fragment.

Track 5 - Deletion of a DNA sequence occurs between the flanking restriction sites, creating a smaller restriction fragment.

Track 6 - One of the flanking restriction sites is lost through mutation or deletion. Consequently, the restriction fragment is lost.

Example - RFLP in *Brassica*



This slide shows RFLP marker segregation in *Brassica* (courtesy of D. Marshall). Each track on the gel was loaded with a separate DNA restriction digest sample. The black bands represent where the radioactive probe hybridised to DNA fragments on the nitrocellulose filter. The difference in banding patterns between the samples reflects genetic differences between the samples.

RFLP: single-locus probes

- Nuclear DNA
  - genomic libraries
  - cDNA libraries
- Cytoplasmic DNA
  - chloroplast and mitochondrial DNA libraries
- Such RFLP probes are:
  - locus specific, codominant
  - mainly species specific

Traditional RFLP analysis makes use of probes obtained from the following sources:

• **Nuclear DNA**; these probes are obtained from:

▪ **genomic libraries** - total plant DNA is digested with restriction enzymes (e.g. *Pst*I) and individual fragments cloned into a bacterial or viral vector. Suitable probes are selected from this "anonymous" library for RFLP analysis.

▪ **cDNA (complementary DNA) libraries** - mRNA is isolated and transcribed into DNA using the enzyme reverse transcriptase. The cDNA so obtained is cloned into vectors and used as a library for probes in RFLP analysis. Also known as copied DNA libraries.

• **Cytoplasmic DNA** - mitochondrial and chloroplast DNA libraries.

In general terms, RFLP probes from these sources are:

- Locus specific - give rise to easily identified codominant markers;
- Mainly species specific.

As a result of the species specificity shown by many single locus probes, it is often necessary to isolate probes from genomic or cDNA libraries for studies on new species. This can be very time consuming. However, in many cases it is possible to use probes from related genera.

RFLP: multi-locus probes

- Tandem repeats are very useful
  - found at numerous loci
  - highly polymorphic
- Minisatellite sequences
  - VNTR: variable number of tandem repeats
  - use in DNA fingerprinting
  - use of M13 bacteriophage repeat sequences

As mentioned earlier, tandem repeats have received considerable attention as molecular markers since they vary according to the size and sequence of the repeat unit, the number of repeats found and their distribution throughout the genome. The repetitive sequences which have particular application are:

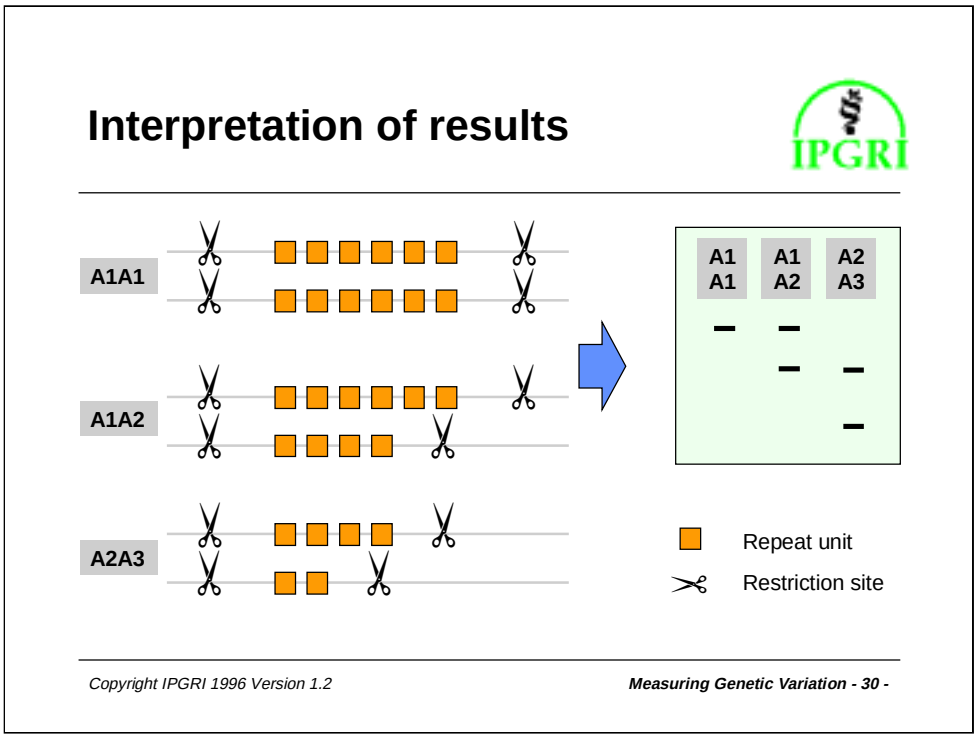
• **minisatellite** sequences - tandem (i.e. head to tail) repeats of a basic 'motif' of about 10 to 60 base pairs which occur at many loci on the genome;

• **microsatellite** or simple sequence repeats (SSR) - tandem repeats of very short motifs of generally 1 to 5 base pairs in length.

Minisatellite sequences

Work on plant minisatellite markers resulted from the pioneering studies on the human genome (Jeffreys *et al.* 1985a, b). Jeffreys' group studied tandem repeats of sequences ranging from 15 to 75 base pairs and found them to be highly variable between human individuals. Since the polymorphisms were related to the *number* of repeated units, the sequences have the alternative name of **VNTRs** (Variable Number of Tandem Repeats; Nakamura et al. 1987). By carefully selecting a probe it was possible to detect restriction fragments representing a large number of loci. The patterns of the minisatellite-bearing restriction fragments on autoradiographs (the so-called 'DNA fingerprint') allowed a clear discrimination between different human individuals.

In plants and many other organisms, probes from an internal repeat from the protein III gene of the bacteriophage M13 have been used to reveal minisatellite sequences (Ryskov *et al.* 1988; Rogstad *et al.* 1988).



This slide (after Ferreira and Grattapaglia 1996), shows the results of RFLP analysis using a probe derived from the repeat unit of a minisatellite sequence (VNTR). The three genotypes are A1A1, A1A2 and A2A3. Since RFLP markers are codominant, it is possible to detect the alleles A1, A2 and A3 in heterozygotes.

Advantages and disadvantages of RFLP

- Reproducible
- Co-dominant markers
- Simple

- Time consuming
- Expensive
- Use of radioactive probes

Advantages

• Results are highly reproducible between laboratories

• RFLP markers usually show codominant inheritance - heterozygotes can be distinguished from homozygotes

• Discriminating power - can be at the species / population level (single-locus probes), or individual level (multi-locus probes)

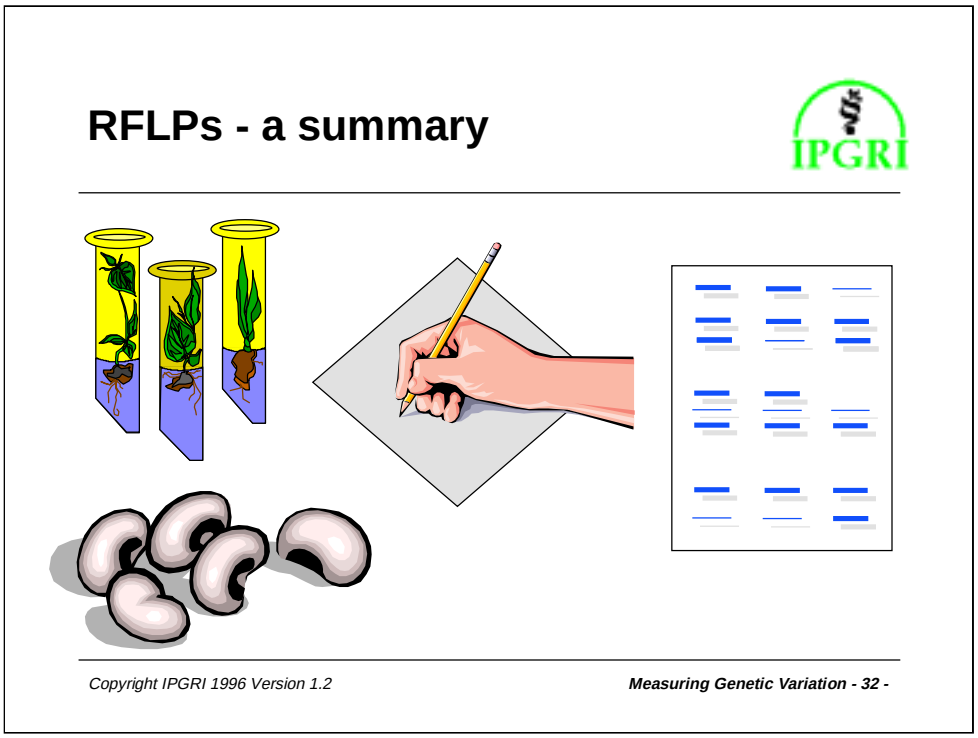
• Simplicity of the method - given the availability of suitable probes, the technique can be applied readily to any system.

Disadvantages

• Time consuming and expensive to perform - technical expertise required

• Where no suitable single-locus probes exist, it is time consuming and expensive to identify suitable marker / restriction enzyme combinations from genomic or cDNA libraries

• Most RFLP work is carried out using radioactively labelled probes and therefore requires expertise in autoradiography. This can be a serious drawback in some situations where special facilities and permits are required to carry out the work.



RFLP analysis involves the study of the variation in DNA restriction fragment length as revealed by hybridisation with a particular probe.

Differences between two genotypes may result from:

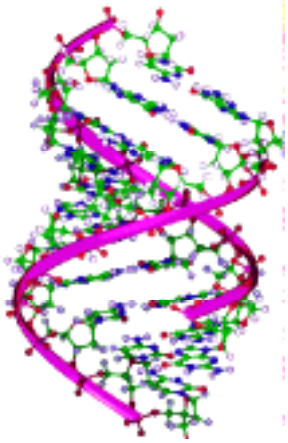
- **point mutations** creating or destroying restriction sites - this affects band presence or absence;
- **reorganisations of blocks of DNA** between two restriction sites, such as deletions or insertions, results in a change in band size.

Single-copy genomic and cDNA probes are generally *species specific*. They have to be developed from DNA libraries for new subject material. This can be very labour intensive and time consuming.

Multi-copy, multi-locus probes have been shown to be useful for fingerprinting in a wide range of plant species, e.g. M13 minisatellite probes.

The main advantages of the RFLP methodology are its discriminating power, its reproducibility and its simplicity. The main disadvantages are that it is time consuming and expensive to perform.

DNA based markers



- DNA basics
- RFLP methods
- **PCR-based methods**
- Sequence-tagged sites
- AFLP
- Electrophoretic techniques

Topic 5 - PCR - Polymerase Chain Reaction

- Powerful technique for amplifying DNA
- Involves the cycling of temperatures
  - denaturing step
  - annealing step
  - elongation step
- Amplified DNA separated by gel electrophoresis

The polymerase chain reaction or PCR (Erlich 1989) is a powerful but simple technique for amplifying a tiny amount of target DNA sequence several million times over in only a few hours. The process is as follows:

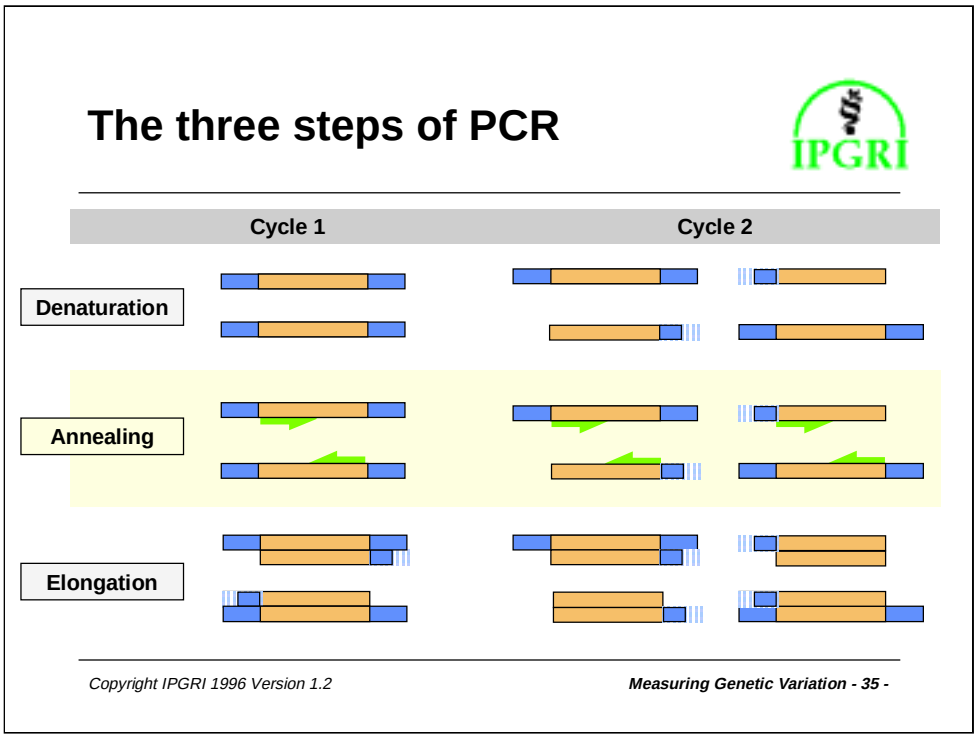
• Suitable DNA (such as total cell DNA following restriction digestion) is mixed with short oligonucleotide primers.

• The following cycle is performed:

- **Denaturing step** - the DNA is denatured at an elevated temperature, reducing it to single strands.
- **Annealing step** - cooling allows the primers to anneal to the complementary region.
- **Elongation step** - the DNA polymerase synthesises a complementary strand.

• This cycle of heating and cooling is repeated many times over, each time the newly synthesised strand becoming a template for subsequent replication. This cycling of temperature results in copying, and then copying of copies, and so on resulting in an exponential increase in the amount of specific copied sequence. Since the amount of DNA placed in the tube at the outset is very small, by far the major DNA component at the end of the reaction cycles is the copied sequence. This is discussed further in the next slide.

• The reaction products are separated by gel electrophoresis. Since they are produced in sufficiently large quantities, they can often be visualised directly by staining with ethidium bromide, thereby avoiding the use of radioisotopes and autoradiography.



In this slide, the target region to be amplified is coloured orange (or light grey), the flanking region is coloured blue (or dark grey) and the primers are indicated by the green arrows.

Cycle 1

In the denaturing step, the DNA strands separate to form single strands. In the annealing step, one primer binds to one DNA strand and another primer binds to the complementary strand. The annealing sites of the primers are chosen so that they will prime DNA synthesis in the intervening region in the elongation step. During the elongation phase, the DNA synthesis proceeds through the target region and for variable distances into the flanking region, giving rise to 'long fragments' of variable lengths (indicated by the vertical bars).

Cycle 2

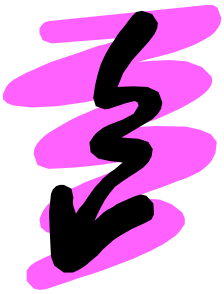
In the second cycle, there are effectively two types of template, namely (1) the original DNA strands; and (2) newly synthesised DNA strands consisting of the target region and variable lengths of the flanking region at the 3' end. When the latter template is used in this cycle, only the target region is replicated.

Cycle 3 (not shown)

In the third cycle, the newly synthesised target region DNA (i.e. without flanking regions) can act as templates. After a few cycles, this DNA fragment quickly establishes itself as the predominant template; repeated cycles therefore lead to exponential increases in the amount of this fragment.

PCR: reaction components

- Use of *Taq* polymerase
- Reaction conditions
- Importance of standardisation



PCR (Erlich 1989) was made possible following the discovery of a new class of enzymes which can copy DNA at elevated temperatures. The first employed was ***Taq* polymerase** (purified from the hot spring bacterium *Thermus aquaticus*).

The PCR reaction contains the following components:

- Target DNA
- Thermostable DNA polymerase
- One or more oligonucleotide primers
- Four deoxynucleotides (dATP, dCTP, dGTP, dTTP)
- Reaction buffer containing the cofactor $MgCl_2$.

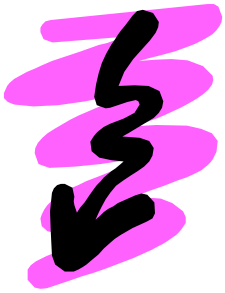
The reaction mixture is taken through the following replication cycles using a commercial thermal cycler:

- **Denaturing step** - one to several minutes at 94-96°C.
- **Annealing step** - one to several minutes at 50-65°C.
- **Elongation step** - one to several minutes at 72°C.

The replication cycles are typically repeated 25 - 50 times. Standardisation of the reaction components and the replication cycles is essential for the reproducibility of results.

PCR - a closer look

- DNA polymerase
- Thermal cycler
- Temperature profile
- Template concentration
- Magnesium ion concentration



• A range of brands and types of **polymerases** are available commercially: *Taq* polymerase (Promega), Ampli*Taq* (Perkin Elmer), Bio*Taq* (Bioline). They vary significantly in price, and different researchers often have their own preferences. However, because different polymerases often give rise to different banding patterns with the same target DNA, the initial choice of enzyme is important, as switching to another enzyme is likely to make comparison with previous results difficult.

• Many brands of **thermal cyclers** are now available, again spanning a broad price range. They vary in features such as the numbers of reactions which can be carried out simultaneously, whether they allow a transition time between different temperatures (ramping), the means and the rapidity of heating and cooling and therefore the time taken for cycling, etc. Despite differences, it is reported that identical fragment profiles can be achieved using different machines.

• Many variations in the **temperature profile** can be found in the literature, and much time can be spent unnecessarily repeating what by now many other researchers have already done to 'optimise' the temperature profile. Originally the temperature profile was as follows (Williams *et al.* 1990): 1 minute at 94°C (for denaturation), 1 minute at 36°C (for annealing), 2 minutes at 72°C (for elongation), 45 cycles.

• Optimisation of **template concentration** is extremely important: a template DNA range of between 5 and 500 ng can be tested. With higher concentrations, impurities can become a problem and this is particularly the case with dried leaf material. In all cases, it is necessary to maintain a standard concentration to achieve consistent results.

• **Magnesium ion concentration** may have a qualitative effect upon banding patterns and therefore will need standardisation.

Topic 6 - RAPD - Random Amplified Polymorphic DNA

- Amplifies anonymous stretches of DNA using arbitrary primers
- Fast method for detecting polymorphisms
- Dominant markers
- Reproducibility problems

The random amplified polymorphic DNA (RAPD) technique is a PCR-based method which uses one or sometimes two short arbitrary primers (usually 8-10 bases) to amplify anonymous stretches of DNA which are then separated and visualised by gel electrophoresis. The key point about this technique is that *nothing* is known about the identity of the amplification products. The amplification products are, however, extremely useful as markers in genetic diversity studies. Other important features of the technique are:

• **The number of fragments**. Many different fragments are normally amplified using each single primer, and the technique has therefore proved a fast method for detecting polymorphisms. The majority of commercially produced primers result in 6-12 fragments; some primers may fail to give any amplification fragments from some material.

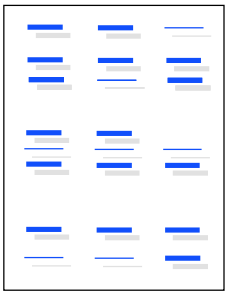
• **Simplicity of the technique**. RAPD analysis does not involve hybridisation/autoradiography or high technical expertise. Only tiny quantities of target DNA are required. Arbitrary primers can be purchased. Unit costs per assay are low. This has made RAPD analysis very popular.

• **RAPD markers are dominant**. Amplification either occurs at a locus or it does not, leading to scores of band presence/absence; this means that homozygotes and heterozygotes *cannot* be distinguished.

• **Problems of reproducibility** - RAPD does suffer from a sensitivity to changes in PCR conditions resulting in changes to some of the amplified fragments. Reproducible results can be obtained if care is taken to standardise the conditions used (Munthali *et al*. 1992; Lowe *et al*. 1996).

Interpretation of RAPDs

- RAPD markers are anonymous
- RAPD markers are dominant
- Co-migration
  - same band, same fragment?
  - one band, one fragment?



As mentioned previously, a key point about the RAPD technique is that since arbitrary primers are used, *nothing* is known about the identity of the amplification products unless the studies are supported by pedigree analysis. This clearly has implications when interpreting band patterns and using the data for statistical analysis. Some considerations are:

• **RAPD markers are dominant.** Homozygotes and heterozygotes cannot be distinguished: RAPD fragments are either present or absent.

• **Problems of co-migration: same band, same DNA fragment?** The presence of a band of identical molecular weight in different individuals is not evidence, *per se*, that the individuals share the same (homologous) DNA fragment.

• **Problems of co-migration: one band, one fragment?** A single band on a gel can be comprised of different amplification products. This is because the type of gel electrophoresis used, while able to separate DNA *quantitatively* (i.e. according to size), cannot separate equal-sized fragments *qualitatively* (i.e. according to base sequence).

PCR with arbitrary primers:
an accumulation of acronyms

- **RAPD**
  - random amplified polymorphic DNA
- **DAF**
  - DNA amplification fingerprinting
- **AP-PCR**
  - arbitrarily primed polymerase chain reaction
- **MAAP**
  - multiple arbitrary amplicon profiling (suggested to encompass all of these minor variations in technique)

All of the following techniques use one or two, short, GC-rich primers of arbitrary sequence. RAPD was the first to become available (Williams *et al.* 1990) and is by far the most commonly used of these techniques.

DAF - DNA amplication fingerprinting

Differences between DAF (Caetano-Anolles *et al.* 1991a,b) and RAPD:

- higher primer concentrations in DAF
- shorter primers used in DAF (5-8 nucleotides)
- two-temperature cycle in DAF compared to 3-temperature cycle in RAPD
- DAF usually produces very complex banding patterns.

AP-PCR - arbitrarily primed polymerase chain reaction

Differences between AP-PCR (Welsh and McClelland 1990) and RAPD:

- in AP-PCR the amplification is in three parts, each with its own stringency and concentrations of constituents
- high primer concentrations are used in the first PCR cycles
- primers of variable length, and often designed for other purposes are arbitrarily chosen for use (e.g. M13 universal sequencing primer).

MAAP is only an acronym proposed by Caetano-Anolles *et al.* (1992) to encompass all of these closely related techniques, but which is not commonly used.

RAPD - a summary

- Fast
- Simple
- Inexpensive
- No use of radioisotopes

- Dominant markers
- Reproducibility problems
- Problems of interpretation

The RAPD technique involves the amplification of anonymous stretches of DNA with one (or sometimes two) short arbitrary primers and the subsequent visualisation of the amplification products by agarose gel electrophoresis.

Advantages

• Fast method for detecting polymorphisms

• Simple - not technically demanding

• Relatively inexpensive to perform - low unit costs

• Avoids the need for hybridisation with radioactive probes.

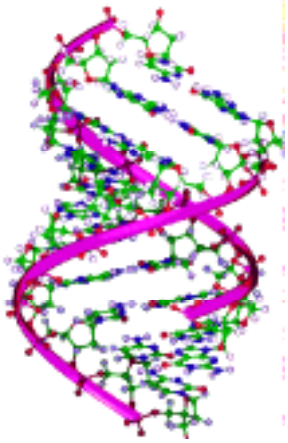
Disadvantages

• Dominant markers - homozygotes and heterozygotes cannot be distinguished

• Problems with reproducibility - RAPDs are sensitive to alterations in PCR conditions

• Problems with interpreting band patterns, e.g. problems of co-migration.

DNA-based markers



- DNA basics
- RFLP methods
- PCR based methods
- **Sequence-tagged sites**
- AFLP
- Electrophoretic techniques

Topic 7 - Sequence-tagged sites

- Sequence-Tagged Microsatellites (STMS)
- Anchored microsatellite oligonucleotides
  - inter-simple sequence repeat (ISSR) primers
- Sequence-characterised amplified regions (SCARs)
- Cleaved amplified polymorphic sequence (CAPS)

More and more sequence information is becoming available from different sources and can be located in widely available databases. This information is extremely useful for developing new strategies for the analysis of genetic variation. A **sequence-tagged site** (STS) is the general term given to a marker which is defined by its primer sequences (Olsen *et al.* 1989). STSs have been used extensively for mapping of the human genome. Examples of STSs are given in the following slides, namely:

- Sequence-tagged microsatellites (STMS)
  - also known as Simple Sequence Repeat Polymorphisms (SSRP)
- Anchored microsatellite oligonucleotides
  - including inter-simple sequence repeat (ISSR) primers
- Sequence-characterised amplified regions (SCARs)
- Cleaved amplified polymorphic sequence (CAPS)

Sequence-Tagged Microsatellites

- Sequence-Tagged Microsatellites (STMS)
  - also known as Simple Sequence Repeat Polymorphisms (SSRP)
- Usually single locus, multi-allelic
- Codominant
- Highly reproducible

Sequence-tagged microsatellites (STMS) - also known as Simple Sequence Repeat Polymorphisms (SSRP) - are an example of STSs. Primers can be constructed which are complementary to the short, unique sequences flanking microsatellite repeat sequence loci and which, in the PCR reaction, direct the amplification of the repeat. Since the repeat length is highly variable, this is an effective way of detecting polymorphisms. These markers generally have the following properties which make them useful for population studies:

- usually define a single, multi-allelic locus
- codominant - homozygotes and heterozygotes can be distinguished
- highly reproducible results obtained.

For high levels of discrimination, polyacrylamide gels are used which can detect single copy differences. It is also possible to combine the PCR reactions with different STMS primers in the same reaction tube (so-called 'multiplexing') which saves on time but this is only possible where the products of the different primers do not overlap in size.

Clearly, the STMS technique requires sequence information for DNA flanking the repeat itself, some of which may be available in DNA databases for well-studied species. Otherwise it is necessary to produce genomic libraries enriched in microsatellites, and from these to select potentially useful clones and then to sequence the DNA in order to define suitable primers (Rassmann *et al.* 1991; Ostrander *et al.* 1992). All of this involves considerable work. Nevertheless, this technique is particularly favoured in population genetics because of the marker's extremely high variability.

Anchored microsatellite oligonucleotides

- Amplification of genomic segments flanked by repeats
- Site specific annealing
- Inter-simple sequence repeats (ISSR)
  - anchored at 3' end
- Markers are usually dominant

Microsatellites are more useful than minisatellites
Core sequences of both satellites can be used

Variants of the STMS technique have been developed using anchored microsatellite oligonucleotides as primers (Zietkiewicz *et al*. 1994) which direct the amplification of genomic segments other than the repeat region itself. These approaches use oligonucleotides based on a simple sequence repeat (SSR) anchored at their 5' or 3' ends by two to four arbitrarily chosen nucleotides which trigger site-specific annealing. This initiates PCR amplification of genomic segments which are flanked by inversely oriented, closely spaced repeat sequences. Specifically, inter-simple sequence repeat (ISSR) primers are anchored at their 3' ends and amplify segments between ISSRs. Such anchored microsatellite markers are usually dominant.

Microsatellite sequences are more useful than minisatellites in these and STMS protocols; many minisatellites are too long to allow amplification using current technology and they are not spread as evenly over the genome as microsatellites. However, the core sequences of both types of satellite sequence may be used.

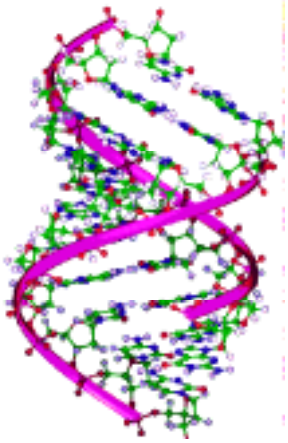
SCARs and CAPS

- SCARs - sequence-characterised amplified regions
  - single-locus marker derived from RAPD fragments
- CAPS - cleaved amplified polymorphic sequence
  - locus-specific marker
  - PCR amplified product analysed by RFLP

Another example of an STS, based on the RAPD technique, is **sequence-characterised amplified regions** (SCARs). These markers are generated by cloning and sequencing RAPD fragments which are of particular interest (perhaps because they are linked to a gene of interest). When the sequence is known, it is then possible to design primers which are longer than the usual RAPD primers (24-mer oligonucleotides) and which are exactly complementary to the ends of the original RAPD fragment. When these primers are used in a PCR, single loci are identified which correspond to the original fragment. These loci are called SCARs. SCARs offer several advantages over RAPD and other arbitrarily primed methods, principally that the results are highly reproducible (longer primers used) and the markers are codominant.

In another technique called **cleaved amplified polymorphic sequence** (CAPS) or PCR-RFLP, PCR primers are constructed for a particular locus. The PCR amplified product is digested with a restriction enzyme and visualised on an agarose gel using ethidium bromide staining. As with RFLP, polymorphisms are detected by differences in restriction fragment sizes.

DNA-based markers



- DNA basics
- RFLP methods
- PCR based methods
- Sequence-tagged sites
- **AFLP**
- Electrophoretic techniques

Topic 8 - AFLP - Amplified Fragment Length Polymorphism

- A combination of RFLP and PCR
- Results in highly informative fingerprints
- An increasingly popular method

Amplified Fragment Length Polymorphism (AFLP) analysis is a highly sensitive method for detecting polymorphisms throughout the genome and is becoming increasingly popular. It is essentially a combination of RFLP and PCR amplification: following restriction enzyme digestion of DNA, a subset of fragments representing many loci is selected for PCR amplification and subsequent visualisation. The technique is also called 'selective restriction fragment amplification' (Zabeau 1993; Vos *et al.* 1995). The procedure is as follows:

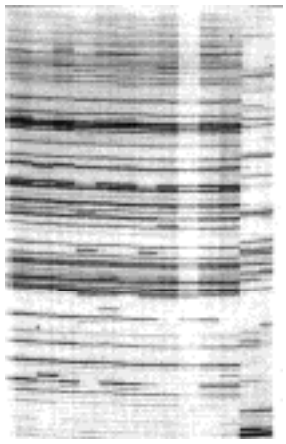
• Genomic DNA is digested with a restriction enzyme.

• Commercial 'adaptors' with defined sequence are ligated to both ends of all the restriction fragments.

• PCR is carried out with specific, commercially produced primers exactly complementary to the adaptors, but whose 3' ends are extended for a few, arbitrarily chosen nucleotides into the fragments. As *exact* matching of the 3' end of a primer is necessary for amplification to occur, only those fragments are amplified which can pair exactly with the 3' primer extension. With a one-nucleotide extension, statistically this will occur for one fragment in 16 (2^4) of the total fragments, and for a two-nucleotide extension for one fragment in 256 (2^8), and so on.

• The amplification products are then separated on highly resolving sequencing gels and visualised using autoradiography. Where radio-labelled nucleotides are not used in the PCR step, fluorescent or silver staining techniques can be used to visualise the products.

Example - AFLP gel for rice



Direction of electrophoresis

This slide shows AFLP analysis in rice (courtesy of J. Xu and G. Bryan). Each track on the gel was loaded with a separate AFLP sample. The black bands represent different PCR amplification products. The difference in banding patterns between the samples reflects genetic differences between the samples.

AFLP

- Highly sensitive
- Highly reproducible
- Widely applicable

- Expensive
- Technically demanding
- Uses radioisotopes
- Problems of interpretation

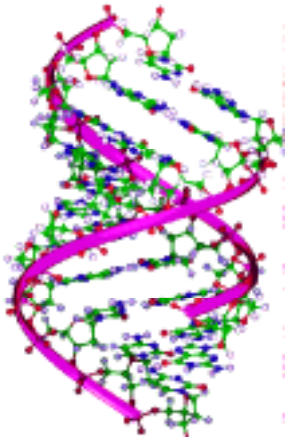
Advantages

- Highly sensitive
- Highly reproducible
- Widely applicable
- Discriminates heterozygotes (when a gel scanner is used)

Disadvantages

- Expensive
- Technically demanding
- Normally uses radioisotopes
- Problems in interpreting banding patterns, e.g. comigration of fragments, uncertainty in assigning equivalence of bands when comparing individuals.

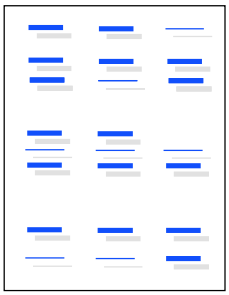
DNA-based markers



- DNA basics
- RFLP methods
- PCR based methods
- Sequence-tagged sites
- AFLP
- Electrophoretic techniques

Topic 9 - Electrophoretic techniques

- DNA sequencing
  - Dideoxy sequencing
- DGGE
- TGGE
- SSCP
- Heteroduplex analysis



DNA sequencing

- The most direct way of studying DNA polymorphism
- Can be highly reproducible and informative
- Expensive
- PCR can be used to target and amplify regions
- Relatively simple if automated sequencing of PCR-derived fragments of DNA is undertaken

DNA sequencing is the most direct way of studying DNA polymorphism: precise information is obtained about the nucleotide sequence of a particular region.

DNA sequencing is most commonly used for studying phylogenetic relationships, where certain regions of the genome are often targeted, such as the *rbc*L gene from the chloroplast (Clegg, 1993), the 18S ribosomal RNA gene (Hamby and Zimmer, 1992), or internal transcribed spacers from the ribosomal genes (Baldwin 1992).

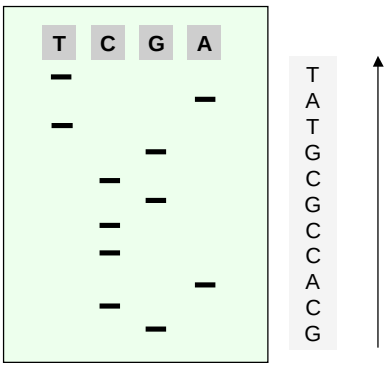
DNA sequencing is difficult, expensive and often involves using radioisotopes for fragment detection. However, the results are highly reproducible and informative.

DNA sequencing is greatly facilitated by the use of PCR for targeting particular regions of DNA and amplifying them in sufficient quantity for sequencing. If automated sequencing machines are available for use, then the process is even simpler, and a lot of information can be gathered in a short space of time, but this is still expensive.

In the future, fluorescence-based automated sequencing may allow sequence information to be obtained routinely from long stretches of DNA, and for sequence data to be recorded by computer for direct analysis to detect polymorphisms.

Dideoxy sequencing

- Selective termination of DNA synthesis
- Four reactions
- Separate by denaturing PAGE
- Visualise by autoradiography



The dideoxy sequencing method (also known as the Sanger or chain-termination method) involves replicating DNA to produce fragments of various lengths, separating these fragments by electrophoresis and then deducing the DNA sequence from the resulting pattern.

This method is made possible owing to the effects of dideoxynucleotide triphosphates in DNA synthesis: these can be incorporated like deoxynucleotide triphosphates but terminate the growing chain wherever they are incorporated. Therefore, in a DNA synthesis reaction where dideoxyadenosine triphosphate (didATP) is present in addition to the normal deoxyadenosine triphosphate (dATP), a series of DNA fragments of varying lengths will be produced, each terminating with a didATP residue at the 3' end. This should occur for each adenosine position on the template. If separate reactions are performed with each of the other dideoxynucleotides triphosphates didTTP, didCTP, didGTP (but using the same template and primer), it is possible to have a series of DNA fragments where DNA synthesis has stopped at each position on the template. These fragments can be separated by denaturing polyacrylamide gel electrophoresis and, if a radiolabelled nucleotide is used, the fragments can be visualised by autoradiography. The sequence can then be read from the bottom of the gel to the top (see slide). In this example, the sequence is GCACCGCGTAT.

Quantitative vs. qualitative separation

- Quantitative separation - by size
  - agarose gel electrophoresis
  - polyacrilamide gel electrophoresis
- Qualitative separation - by sequence
  - sophisticated equipment required
- DGGE - Denaturing Gradient Gel Electrophoresis

DNA profiling or fingerprinting relies on gel electrophoresis to separate the DNA fragments according to their sizes followed by staining or autoradiography to visualise the banding patterns. Owing to its simplicity, agarose gel electrophoresis is commonly used. Where greater resolution is required, such as with low molecular weight fragments, polyacrylamide gels are used. Such one-dimensional separation methods, while able to separate DNA *quantitatively* (i.e. according to size), cannot separate equal sized fragments *qualitatively* (i.e. according to base sequence). DNA sequencing in such situations, although accurate, is not a practical option for detecting sequence differences in large-scale screening. Accordingly a number of electrophoretic methods have been developed which address this problem, many requiring sophisticated equipment.

DGGE - Denaturing Gradient Gel Electrophoresis

In this method, small DNA fragments (usually less than 1 kb) are electrophoresed through a polyacrylamide gel under increasingly denaturing conditions (usually increasing formamide/urea concentrations) until the DNA 'melts' and becomes single-stranded. At this point there is a sharp reduction in mobility owing to changes in shape and the fragment virtually stops. The point at which the DNA melts is dependent on the nucleotide sequence in the melted region; sequence differences between different fragments can therefore cause the DNA to melt at different positions on the gel. DGGE can be used in a two-step electrophoresis procedure to fingerprint a sample. Firstly, the sample is electrophoresed to separate to fragments according to size. Secondly, it is electrophoresed under DGGE conditions to separate according to the 'melting point'. When samples are mixed, double-bands indicate sequence differences between bands of the same size.

TGGE, SSCP and Heteroduplex analysis

- TGGE - Thermal Gel Gradient Electrophoresis
- SSCP - Single-Stranded Conformational Polymorphism
- Heteroduplex analysis

TGGE - Thermal Gel Gradient Electrophoresis

This method is similar in principle to DGGE except that the DNA is denatured across a temperature gradient. The technique can also be used for analysing single-stranded RNA and proteins (Riesner *et al*. 1991).

SSCP - Single-Stranded Conformational Polymorphism

This method of separation relies on the differential mobility of single stranded DNA on polyacrylamide gels. Single-stranded DNA forms a secondary structure which is unique to the particular sequence. The differences in secondary structure cause the DNA strands to migrate differentially on the gel, each strand ideally having a unique mobility (Hayashi 1992).

Heteroduplex analysis

In this technique, two PCR-amplified products are mixed in equal quantities, denatured at 95°C and allowed to cool down. During the cooling process, DNA strands from the different PCR products will re-anneal to form heteroduplex DNA. Any mismatches in the heteroduplex DNA will cause it to have a different three-dimensional structure to the homoduplex DNA and therefore a lower mobility on polyacrylamide gels. The reduced mobility is proportional to the degree of divergence of the sequences (Delwart *et al*. 1993).

Topic 10 - Considerations when choosing a technique

- What information is required?
- At which level is discrimination sought?
- How many loci are required?
- Reproducibility of results
- Cost
- Speed

• **What information is required?** It is important to choose the right technique in order to obtain the right information. For instance, for information on population history or phylogenetic relationships, sequence data or restriction site data should be obtained.

• **Discrimination -** at which taxonomic level is the genetic variation being measured - within populations, between species, between genera? Do the methods detect the desired level of variation?

• **The number of loci** - will information from a few or many loci be required? Seed proteins yield information on only a few loci, and allozymes are limited. RFLPs and RAPDs provide high numbers of loci.

• **Reproducibility** - Are robust, reproducible methods required? If so, allozymes, RFLPs, STMS and sequencing are robust, while RAPD is not.

• **Cost** - allozymes are the least expensive, RAPD and RFLP intermediate, AFLP and sequencing most expensive. If probes or primers are not available for techniques such as STMS and SCARS, significant costs can be incurred developing them.

• **Speed** - PCR-based methods certainly give rapid results where primers are available. Hybridisation-based methods are slower. Conventional DNA sequencing is slow but automated sequencing is faster.

Considerations (cont$^{d.}$)

- Is the right expertise available?
- How much good-quality DNA is available?
- Is the mode of marker inheritance important?

• **Expertise required** - Techniques involving hybridisation / autoradiography or sequencing are technically demanding. RAPDs are the least demanding and unit costs per assay are low.

• **DNA availability** - Most PCR-based methods require only tiny quantities of easily prepared DNA, RFLP analysis requires larger amounts, and sequencing requires the greatest quantities.

• **Mode of inheritance** - Is it necessary to identify homozygotes and heterozygotes? Are codominant markers needed (single-locus RFLPs, allozymes, PCR-amplified microsatellites), or will dominant markers suffice (RAPD, AFLP)?

• **Number of alleles required** at individual loci - Where hypervariability is required, no other techniques can compete with those based on single locus, simple sequence repeats (e.g. STMS).

Getting started - practical considerations

- Resources
  - well-equipped laboratory
  - money to purchase equipment and consumables
  - laboratory skills
- Importance of PCR-based techniques
- RAPD and microsatellites
- AFLP

How do you decide which DNA molecular marker techniques to embark upon first for studying genetic resources? Clearly this will, in the first instance, depend upon (a) having access to a well-equipped laboratory; (b) having money to purchase additional equipment and consumables when necessary; (c) having a good grasp of many basic laboratory skills. It is generally invaluable to make an extended visit to another laboratory where relevant techniques are being practised before starting. After this, where to go? Which techniques to use?

• PCR-based molecular marker techniques must take priority, as opposed to hybridisation-based techniques which are labour intensive and require special facilities and permits in order to carry out the work. When PCR has been learned, it opens up numerous other possibilities.

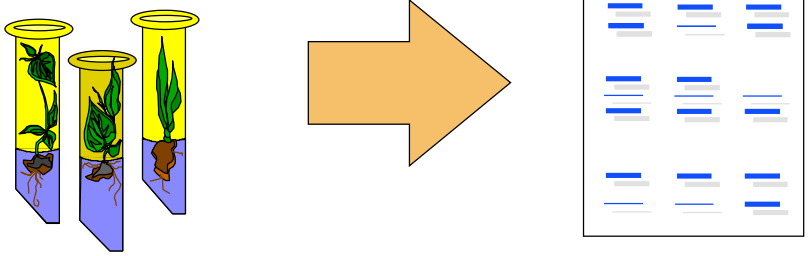
• Despite the general lack of 'portability' of the RAPD technique, it is an excellent technique to begin with. It is ideal for screening for polymorphisms in most, if not all, species of interest, and is particularly attractive because components such as primers are readily available.

• Once the RAPD technique has been mastered, then PCR-based SSR or microsatellites could be developed relatively easily. However, if suitable primers are not available, then the amount of work necessary to develop primers would prohibit the use of this marker technology. Primers are generally available for rice, brassicas, etc.

• AFLP is rapidly becoming a very popular (although expensive) option. Proficiency with the RAPD technique, combined with a visit to an appropriate laboratory would be a good start before using this method.

Topic 11 - Practical applications

- Rice studies
- Other cultivated plants
- Genebanks



Rice studies

- Studying polymorphism in landraces and cultivars of rice
- Identification of crossability groups
- Identification of cultivars
- Diversity studies in rice germplasm
- Rice phylogenetic studies

• Dallas (1988) was able to distinguish rice cultivars using a human minisatellite probe and prove Mendelian inheritance of fragments.

• Yang *et al.* (1994) used PCR technology to identify microsatellite polymorphism across landraces and cultivars of rice.

• Wu and Tanksley (1993) reported the identification of microsatellite alleles that are specific to either *indica* or *japonica* rices.

• Glaszmann (1987, 1988) used isozymes to identify all the crossability groups.

• Virk *et al.* (1995) used RAPD for separating different types of cultivated rice, and for identifying duplicates.

Other cultivated plants

- Studying domestication
- Cultivar identification - taxonomy
- Genetic relatedness/diversity: population genetics
- Pedigree analysis
- Hybrid identification
- Gene mapping and identification
- Marker-assisted selection in plant breeding
- Monitoring and identifying somaclonal variation

• Gepts (1995) provides a useful overview of several studies of domestication using molecular markers, and includes *Zea, Phaseolus* and *Brassica.*

• Nybom *et al.* (1990), Koller *et al.* (1993) used M13 and RAPD respectively to identify apple cultivars.

• Tinker *et al.* (1993), Hu and Quiros (1991), Kresovich *et al.* (1992), Yang and Quiros (1993), Dweikat *et al.* (1993), Wilkie *et al.* (1993) and Mori *et al.* (1993) used RAPD to characterise varieties of barley, brassicas, celery, oat, onion, potato and cotton.

• Dawson *et al.* (1993) studied patterns of variation in *Hordeum spontaneum.*

• Doebley and Stec (1993), Takumi *et al.* (1993) and Wilkie *et al.* (1993) made important comparisons of DNA sequence data, map data and other marker data in phylogenetic studies.

• Wolff *et al.* (1995) studied the stability of DNA fingerprints of *Chrysanthemum* cultivars.

• Salimath *et al.* (1995) looked at genome origins and genetic diversity in *Eleusine.*

• Margale *et al.* (1995) and He *et al.* (1995) studied genetic variation in germplasm collections of French brassicas and sweet potatoes respectively.

Genetic resources and genebanks

- Taxonomic characterisation of germplasm
- Evaluation of germplasm for useful genes
- Maintenance of collections
  - identifying gaps
  - identifying duplicates
  - development of core collections
- Development of conservation strategies

Molecular markers may be employed to assist genebank management, organisation and the way that material is accessed:

• for the accurate identification of germplasm
• the screening of germplasm for use by breeders and other researchers in making crosses, mapping, identifying and isolating genes of interest
• the routine maintenance of germplasm, which will be streamlined by the identification of duplicates and the development of core collections
• the identification of gaps in the collection - useful for planning future conservation and collecting activities

• **Examples of taxonomic identification of germplasm:** The identification of indica and japonica rices (Virk *et al.*, 1995a; Zhang *et al.* 1992; Zhang *et al.* 1994; Wu and Tanksley, 1993). The ease with which this discrimination can be made using RAPD markers is illustrated in the next slide.

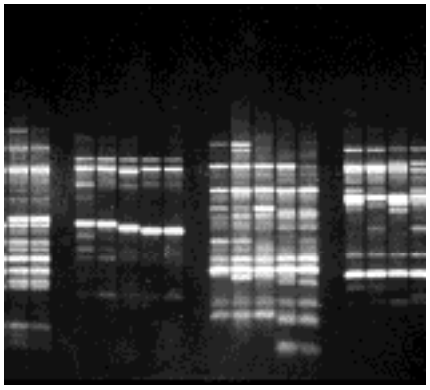
• **Example of duplicate identification:** The identification of true and suspected duplicates in rice germplasm collections (Virk *et al.* 1995b). Two procedures have been proposed for identifying such duplicates.

Core collections and conservation strategies

Molecular markers can help in the choice of core material by ensuring that the allelic richness of a core will be maximised. Schoen and Brown (1995) suggest that genetic marker data be utilized to guide sampling.

In the same way, molecular data can be used to define conservation strategies, both *ex situ* (e.g. collecting strategies) and *in situ*.

Example: RAPDs in rice



Polymorphism for RAPD markers in rice germplasm highlighting diversity: five accessions using three primers (slide courtesy of B. Parsons).

The Future

- Automated methods
- Multi-allelic, locus-specific markers
- PCR-based identification

- **Faster analysis using automated methods**: mass fingerprinting
  - simplified DNA extraction and purification
  - elimination of gel electrophoresis for identifying fragments of DNA
  - use of automated work stations
  - direct data analysis by computer

- **Increased use of multi-allelic, locus-specific markers**

For population and genetic mapping studies, data from techniques using multi-locus markers (e.g. RAPD) can be ambiguous without pedigree analysis. With the increased availability of sequence information, it is possible to design locus-specific primers to detect polymorphisms (e.g. STMS, SCARs and CAPS). The future here looks very bright.

- **PCR-based identification of genes of interest**

Using PCR-based methods, markers can be used for analysing genes of interest within germplasm collections during the evaluation process. It is now becoming known that the linear arrangement of genes on chromosomes has been broadly maintained during evolution across a wide spectrum of biodiversity, so-called synteny (Devos *et al*. 1995). With this knowledge, it is now possible to exploit genetic resources to an increased extent. For instance, by knowing the position of a gene in the genome of one species or genus of grass (say rice), it will enable the location of the same gene in a different species or genus (for example, maize or sorghum).